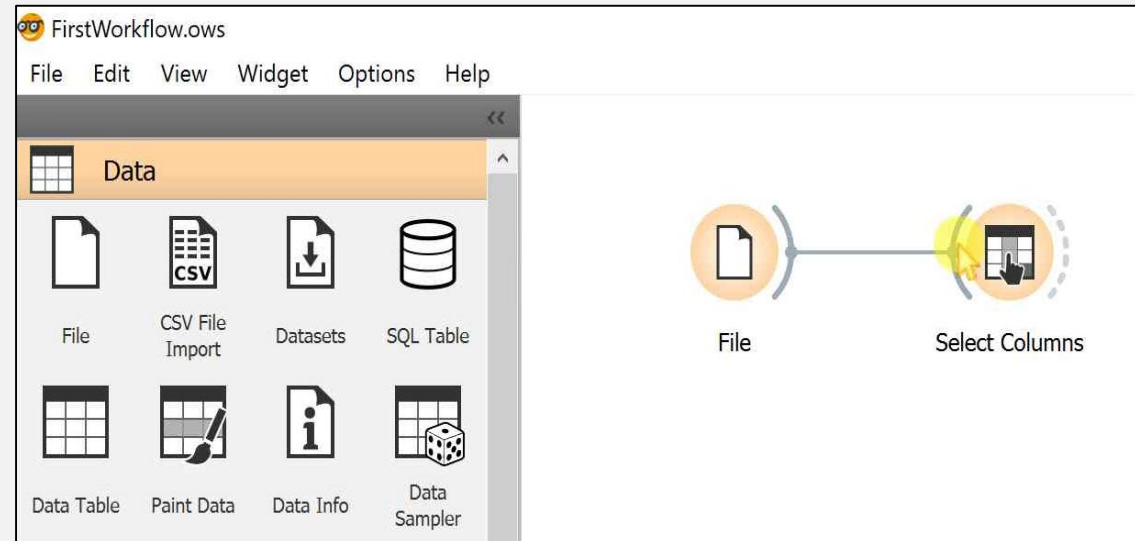
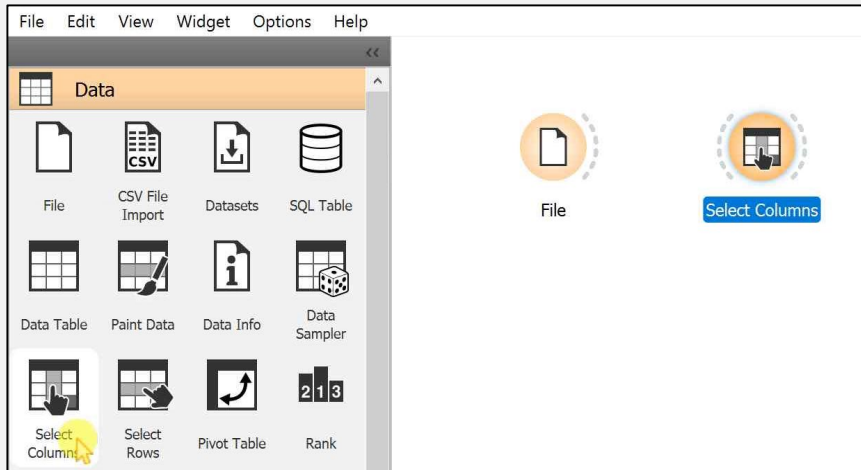


Select Columns

Choose the **Select Columns** widget to add it to your work area.
We will use this widget to set up the fields we want to use in our analyses.

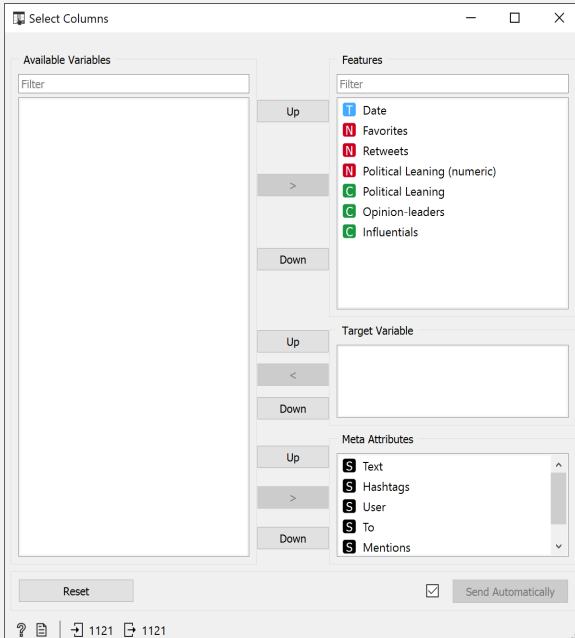


To connect the two widgets, left-click on the output side (right) of the File widget and hold down the mouse button while dragging the connecting line over to the input side (left) of the Select Columns widget.



Select Columns

Double-click on the **Select Columns** widget to open its options screen.

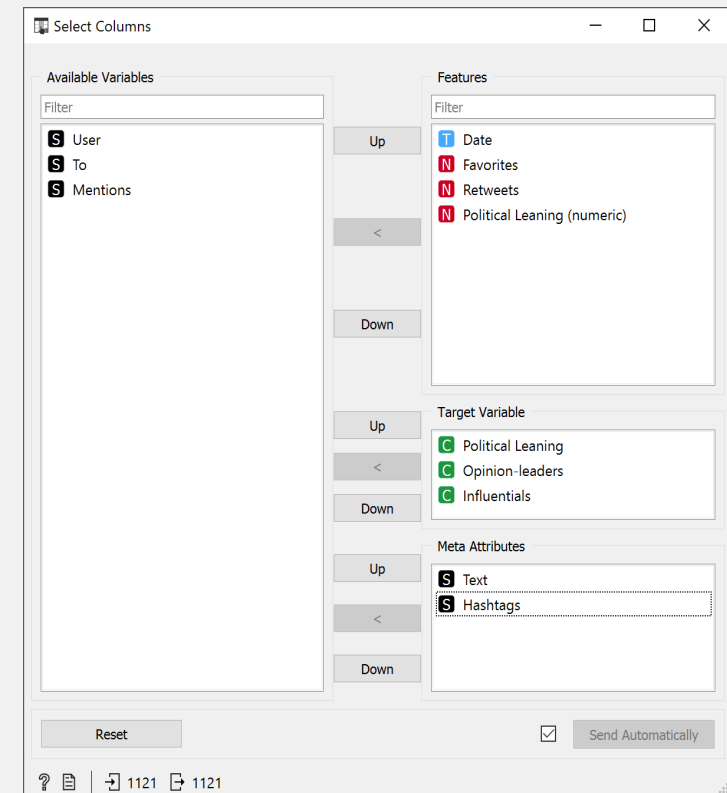


Default

You may **drag and drop** the fields into the different areas of the screen or highlight them and use the arrows in the center section.

For this **first workflow**, since you are using your own spreadsheet of non-categorized data, leave the target variable area empty.

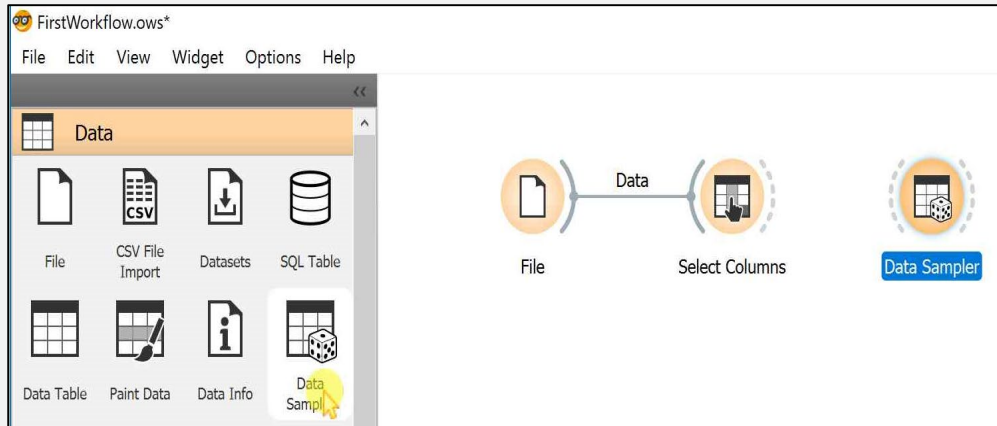
In this tutorial we will not be analyzing the User, To, or Mentions field so those have been moved out of the Meta Attributes area.



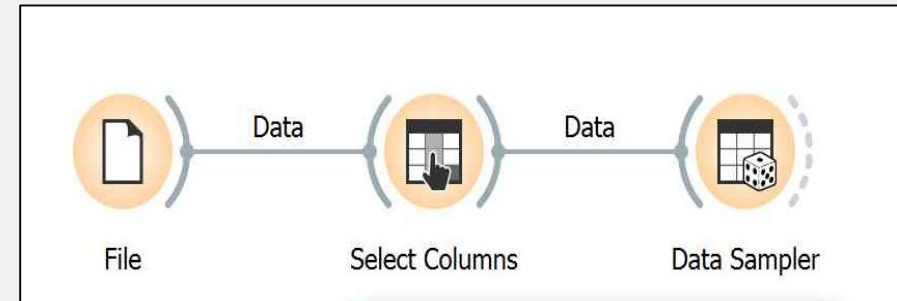
Once the column names have been arranged in the correct areas, you can close this screen.

Data Sampler

If your dataset is quite large, or just to save time while you develop your workflow, you can use the Data Sampler widget to select a small portion of your data.



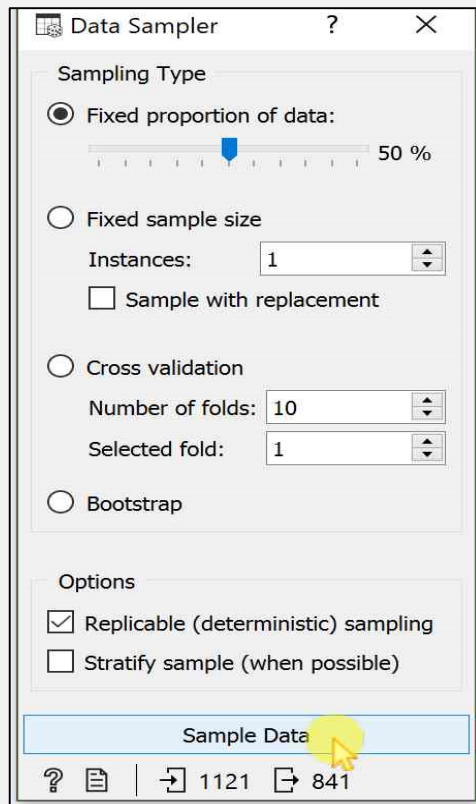
Add a Data Sampler widget to your work area.



Connect the output of the **Select Columns** widget to the input of the **Data Sampler** widget and then **double-click** on the Data Sampler to open the options screen.

Data Sampler

The default for this widget is 75% and if you look at the bottom line in the image below, you will see that the 1,121 rows are reduced to 841. After dragging the slider to 50%, you must click on the **Sample Data** button at the bottom to update the sample.



For this tutorial, we are going to reduce the size of the sample dataset to speed up our initial setup of the workflow.

Since you are using **your own dataset** for this portion of the tutorial, use the **Fixed proportion of data** slider and **reduce your data size down to less than 1,000 rows**.

Once you have sampled your data, you can close this screen



Corpus

Most of the text analysis widgets work with a Corpus of texts. Up to this point in our workflow, each of the widgets takes in, and outputs “data.” Now that our data has been imported and sampled, we need to convert it to a corpus for further processing.



The Corpus widget is contained in the **Text Mining** section of the widget toolbox.

If you **scroll down** to the bottom of the widget window, you will see various other sections of widgets.

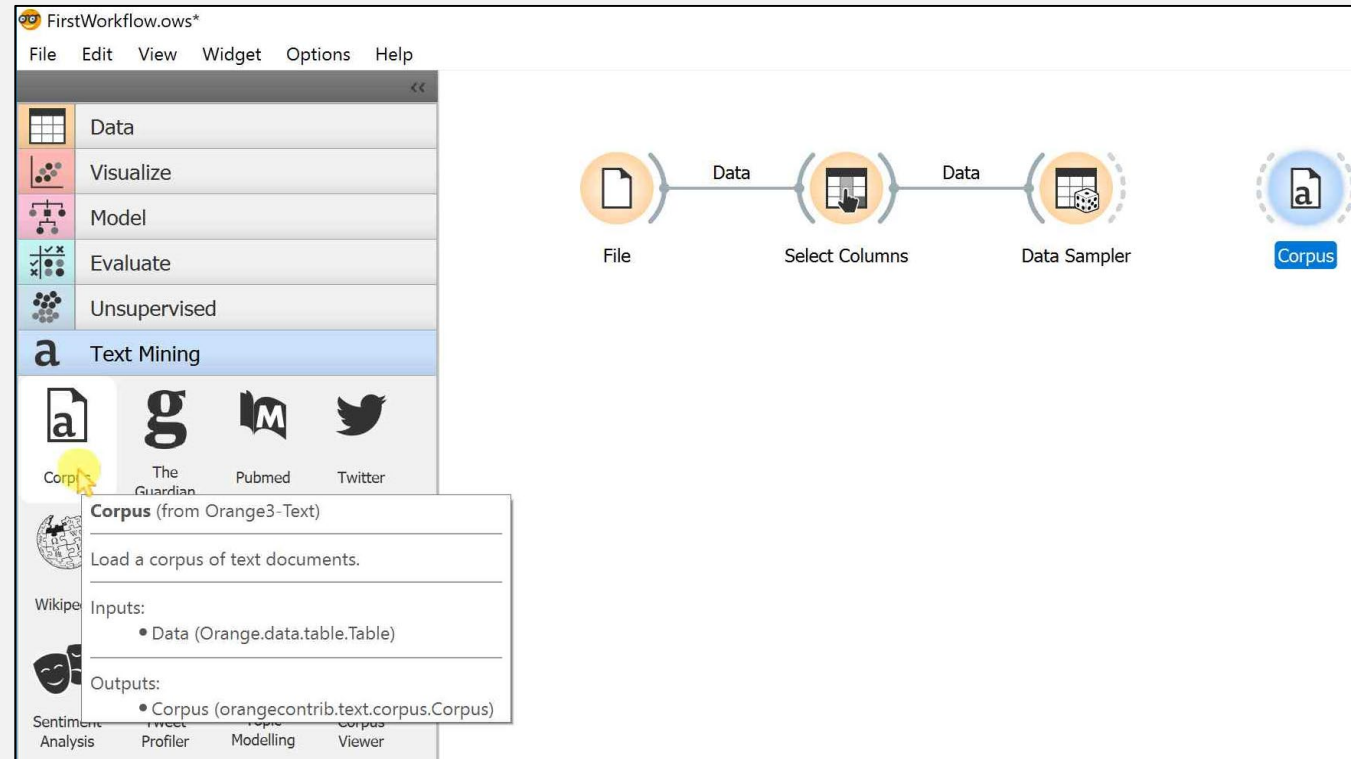
We will use some of the other sections such as Visualize later but for now, click on **Text Mining** to open that section.

Corpus

Click on **Corpus** to add the widget to your work area

When you hover your mouse over a widget in either the toolbox or the work area, it will display a brief indication of the inputs and outputs of the widget.

As seen in this image, the Corpus widget takes Data as its input and sends a Corpus out.



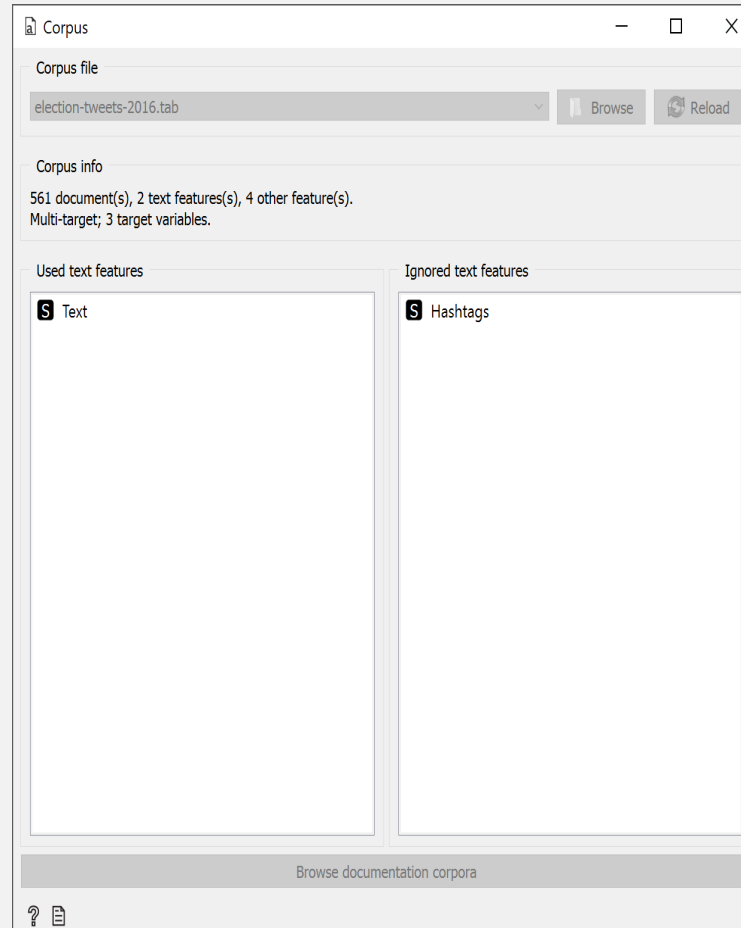
Corpus

Connect the output of the **Data Sampler** widget to the **Corpus** widget. Then **double-click** on the Corpus widget to open the options screen.

This widget screen can be confusing as it is made to be used in two ways. One way is to open a file that is already formatted as a corpus and the other way to use the widget is as we are doing now—converting data to a corpus.

The top section of the screen is greyed out as it is not available when the widget has input from a data source.

So be sure to connect this to your **Data Sampler** first and ignore the filename at the top.



Drag the Hashtags field from the Ignored text features to the Used text features. Once both fields are in the left column, close this window.



Corpus Viewer

The Corpus View widget allows you to look at your data within Orange.

Add the widget to your work area and connect it to the Corpus widget and **double-click** to open it.

The screenshot shows the 'Corpus Viewer' window with the following components:

- Info Panel:** Documents: 561, Preprocessed: False, Tokens: n/a, Types: n/a, POS tagged: False, N-grams range: 1-1, Matching: 270/561.
- Search features:** A list of features with checkboxes. 'Political Leaning (numeric)' is selected and highlighted in yellow.
- Display features:** A list of features with checkboxes. 'Date', 'Favorites', 'Retweets', and 'Political Leaning (numeric)' are selected.
- RegExp Filter:** A text field containing the value 'left'.
- Document List:** A list of 16 documents, numbered 1 to 16, with 'Document 1' selected.
- Document Details:** A panel showing the details for the selected document (Document 1):
 - Date:** 2020-03-26 00:00:00
 - Favorites:** 226
 - Retweets:** 118
 - Political Leaning (numeric):** 1
 - Political Leaning:** left (highlighted in yellow)
 - Opinion-leaders:** yes
 - Influentials:** yes
 - Text:** Really @GOP @realDonaldTrump ? At 10am today New York State reported 199 deaths and by 6pm they reported 280 deaths That's an average of 10 death per hours & Today Texas Declares State of Disaster Due To COVID-19 Is that what you are calling Success @realDonaldTrump ?
 - Hashtags:** ?

The first column has the features you can use to Search and filter the documents to view.

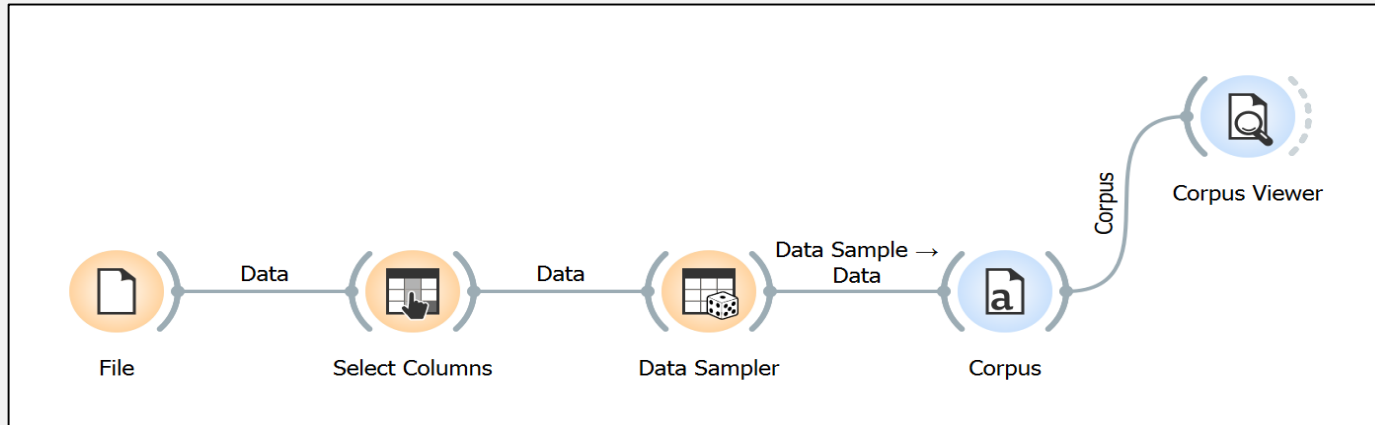
In the image, Political Leaning was selected as the search field and “left” entered in the top RegExp Filter field.

The documents now listed are the tweets categorized as having users that are left leaning.



Workflow Check

At this point in the tutorial, your workflow should resemble the one in this image.



You can drag and drop the widgets anywhere on the work area.

Drag the Corpus Viewer above the current line of widgets as we want to make room for our text analysis items.

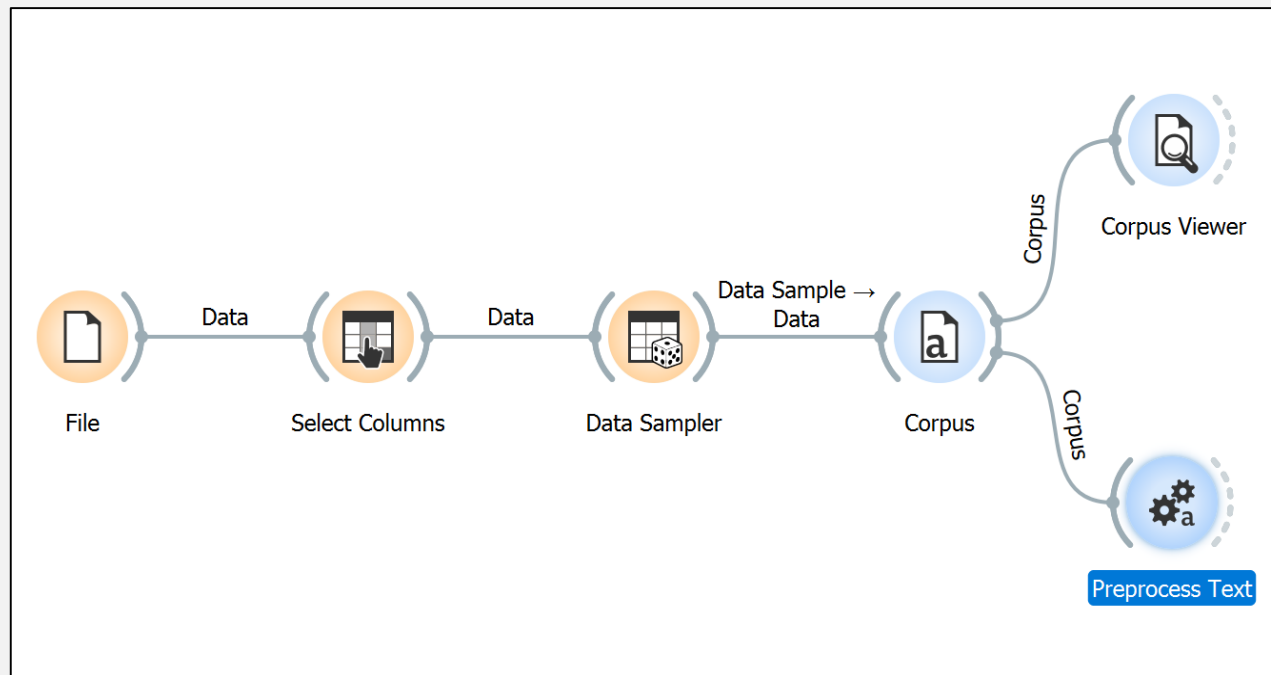
We have imported our data file; chosen the columns for features, target variables, and meta variables; sampled our dataset down to less than 1,000 rows; turned that smaller dataset into a corpus; and viewed and filtered the corpus.

We are now ready for text processing.

Preprocess Text

The output from a widget can be sent to multiple widgets. This saves us from having to create separate workflows that duplicate steps.

Add a Preprocess Text widget to your work area and **connect** it to the Corpus widget.



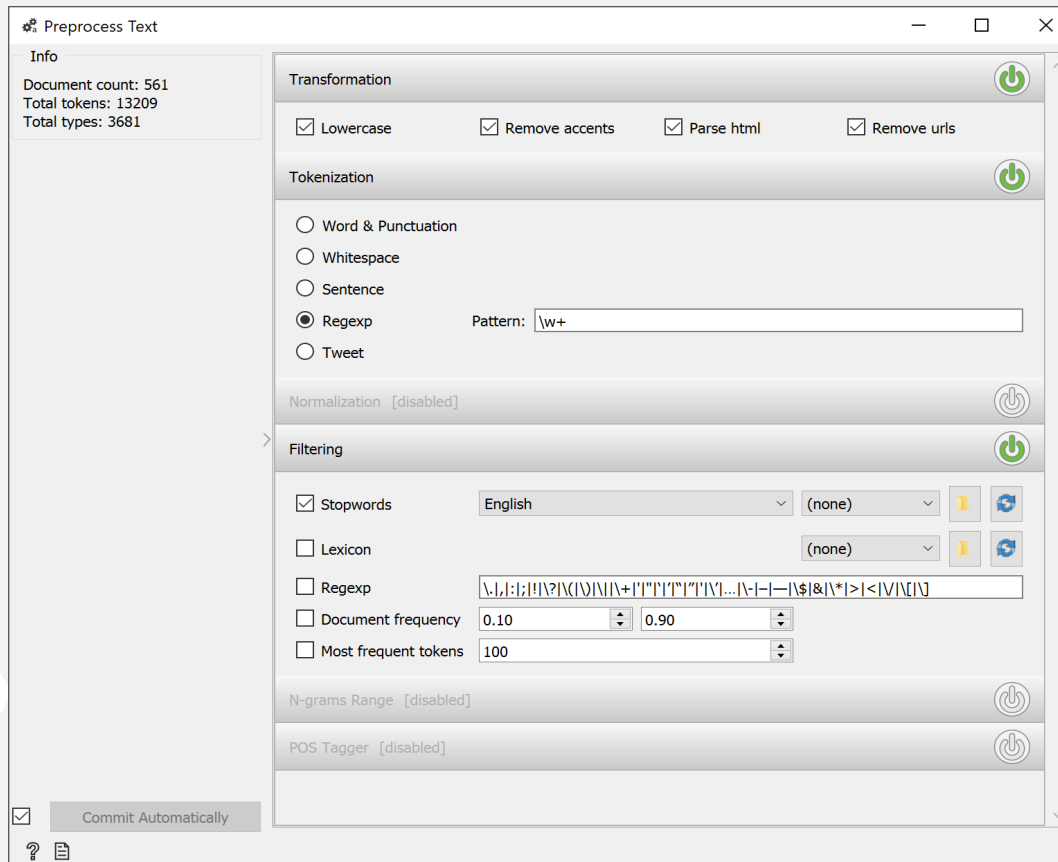
Widgets are capable of one-to-many relationships.



Preprocess Text

Double-click the Preprocess Text widget to open the options screen.

The green on/off buttons activate sections of this screen. By default, the Transformation, Tokenization, and Filtering options are on. The preprocessing splits the text (in our case tweets) into tokens that are used in further analysis.



By checking the boxes in the Transformations section it allows you to remove HTML tags, URLs, and accents from characters (i.e. it uses **e** instead of **è**).

The Tokenization section is set to use regular expressions to break the tweets apart by word.

The Filtering section has Stopwords checked and the language set to English. **If your tweets are in a language other than English, you should switch the language.**

Make your selections match the sample screen above and close this window.

Word Cloud

The last widget we will add to this first workflow is the Word Cloud.

Add the widget to your work area, **connect** it to the Preprocess Text widget and **double-click** on the Word Cloud to open the options screen.

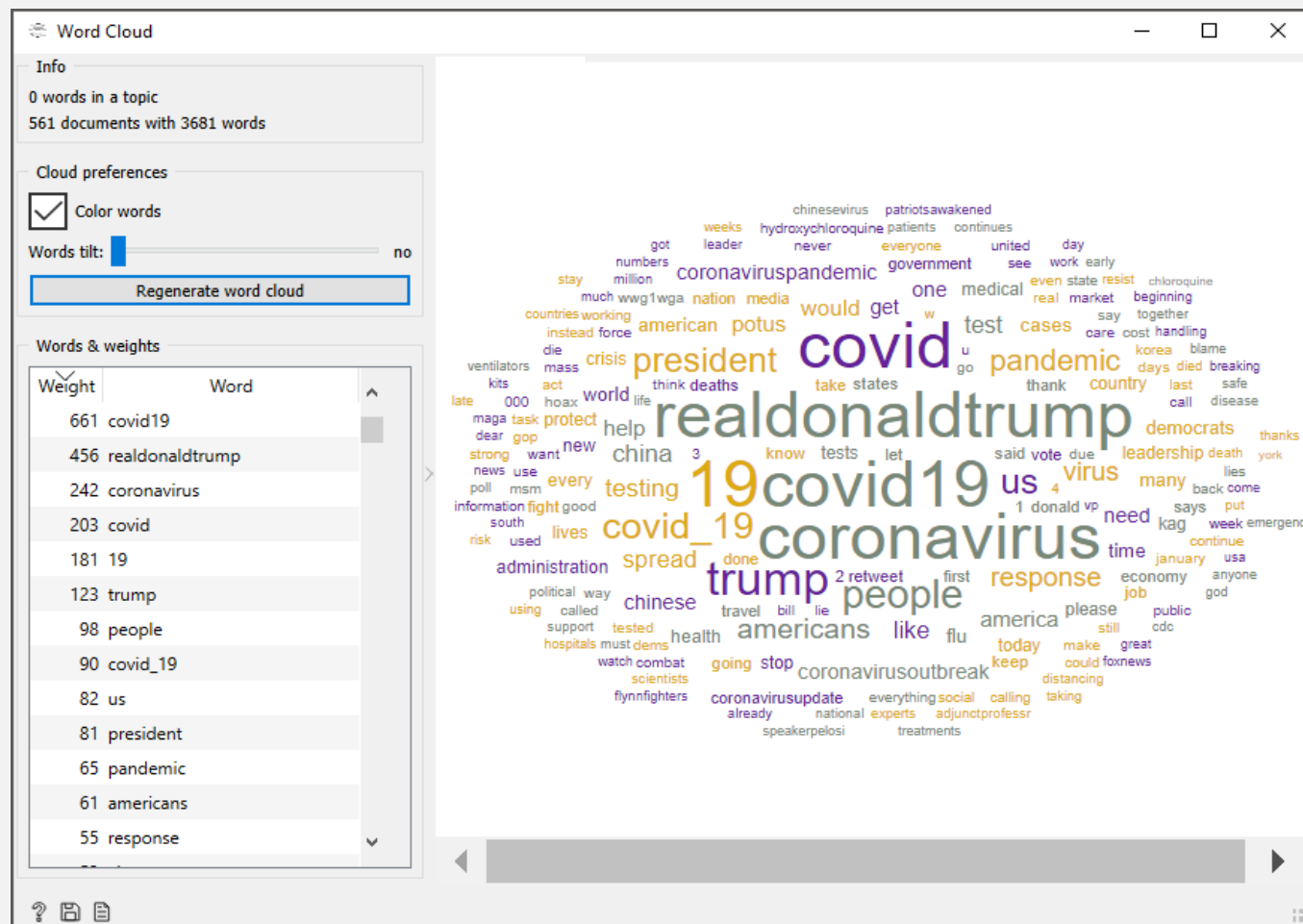
The Words & Weights show the content in tabular form.

Use this to explore your data.

The fourth and fifth items (“covid” and “19”) were split by the tokenization routine on the hyphen from “covid-19”.

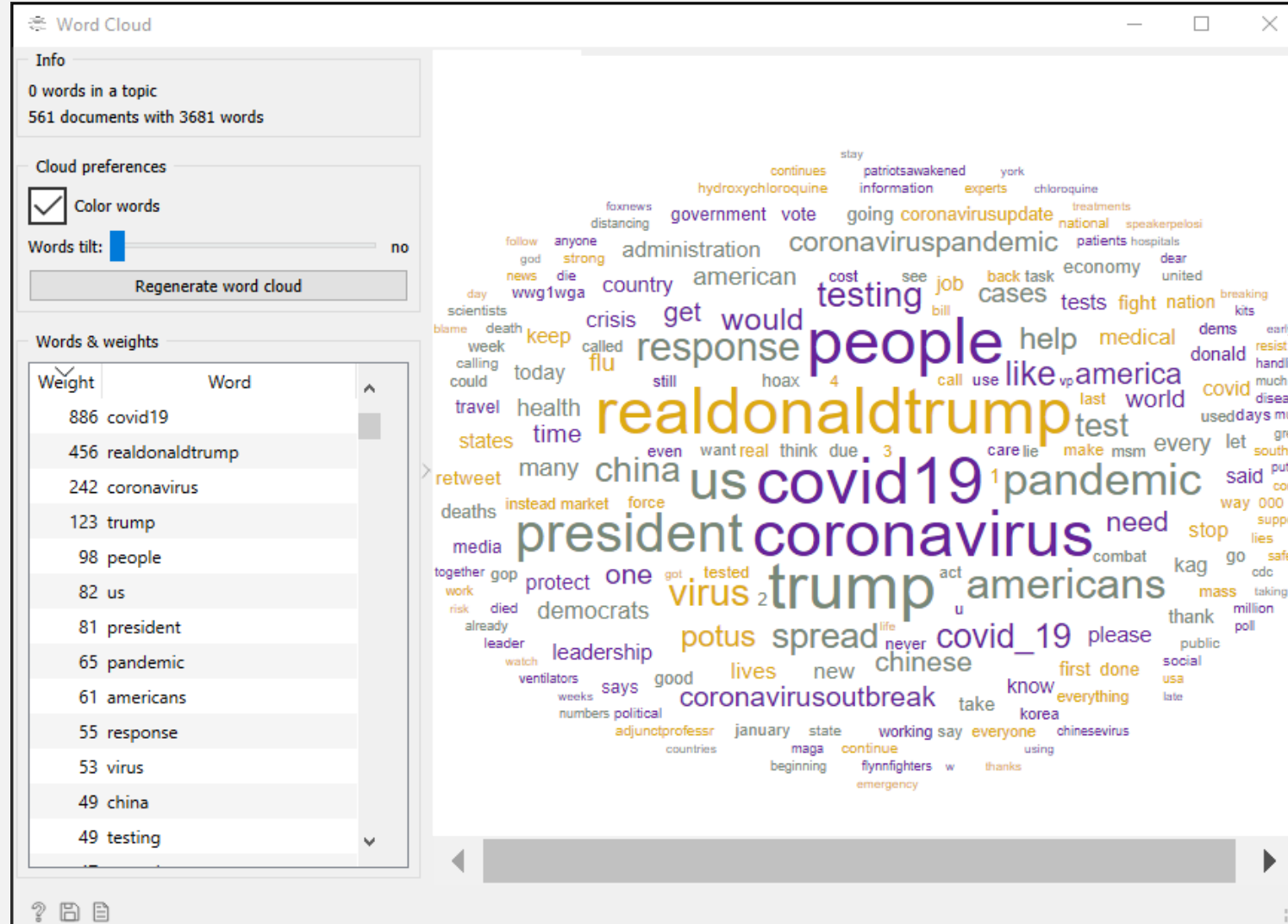
Also notice that covid_19 (with an underscore) is also treated separately.

Inconsistencies in your data can be modified back in the spreadsheet.



Word Cloud

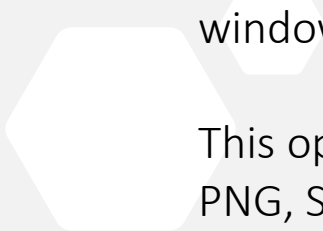
Going back to the spreadsheet data and replacing “covid_19” and “covid-19” with “covid19” in the **tweet text** and then reloading the data in the File widget, the Word Cloud automatically updates.



covid_19 still exists in the word cloud even though the tweet text was updated in the spreadsheet.

This is because we included the hashtags back when we created the Corpus widget.

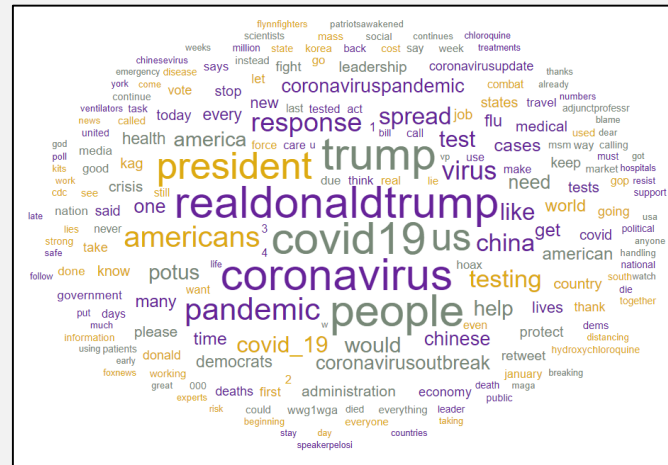




This option allows you to save it as a PNG, SVG, or PDF

For attendees of the NEH Digital Culture Institute, please submit the following as a PDF document:

-
- The diagram shows a sequential workflow for creating a corpus:
- File** (represented by a document icon) is processed into **Data**.
 - The **Data** is then processed into **Data Sample** (represented by a document icon with a magnifying glass).
 - The **Data Sample** is processed into **Corpus** (represented by a document icon with a magnifying glass).
 - The **Corpus** can be used in two ways:
 - It can be processed into **Corpus Viewer** (represented by a document icon with a magnifying glass).
 - It can be processed into **Preprocess Text** (represented by a document icon with a magnifying glass and a gear icon).
 - The **Preprocess Text** is then processed into **Word Cloud** (represented by a document icon with a magnifying glass and a cloud icon).



For Questions:

 Amy Larner Giroux

 407.823.1226

 AmyG@ucf.edu

 CHDR.cah.ucf.edu

